INFORMACIÓN Y ENTROPÍA EN ECONOMÍA

*Álvaro Montenegro**

Desde el punto de vista económico, transmitir, recibir y procesar (entender) un texto o mensaje tiene un costo y exige un esfuerzo que solo se justifica si el beneficio esperado es mayor que el costo. El beneficio depende del nivel de información contenido en el mensaje.

La teoría de la información, también conocida como estadística de la información o teoría de la comunicación, además de estudiar las formas más eficientes de almacenar, codificar, transmitir y procesar la información, se ocupa de la tarea de definir qué es información.

El concepto de entropía, cercano al de información, y asociado con el caos, tuvo origen en la termodinámica, donde describe la variación de la energía no utilizable de un sistema (la cual crece en los sistemas cerrados), y en estadística mecánica, donde el concepto se reformuló para que fuera proporcional al número de estados que puede tomar potencialmente un sistema; en este sentido, es similar a un espacio muestral. En una formulación posterior, la cual se sigue en este artículo, el contenido de información es función del inverso de las probabilidades de los eventos del espacio muestral, y la entropía es el promedio o valor esperado de dicha información.

Para ilustrar esta definición, relacionada con el grado de sorpresa del contenido de un mensaje, veamos algunos ejemplos. Si recibimos el mensaje "mañana el sol se elevará sobre el horizonte", pensaremos que el mensaje no da mucha información y no vale la pena gastar recursos en transmitirlo y procesarlo pues solo reporta el hecho predecible de

* Doctor en economía, profesor titular del Departamento de Economía de la Pontificia Universidad Javeriana, Bogotá, Colombia, [amontene@javeriana.edu.co]. Fecha de recepción: 14 de julio de 2011, fecha de modificación: 27 de octubre de 2011, fecha de aceptación: 28 de octubre de 2011.

que la tierra continuará girando sobre su eje, trayendo el día y la noche tal como ha ocurrido durante millones de años. Pero, "mañana el sol no se elevará sobre el horizonte" transmite algo que no esperábamos, con un alto grado de sorpresa y de información.

Intuimos que cuanto mayor sea la sorpresa de lo que se transmite mayor será la cantidad de información que contiene el mensaje, y viceversa. El grado de sorpresa se puede asociar al concepto de probabilidad matemática; la sorpresa es mayor cuanto menor sea la probabilidad de ocurrencia del evento reportado, y viceversa. Específicamente, la cantidad de información es inversamente proporcional a la probabilidad de ocurrencia del evento en cuestión. Si es cercana a 1, es decir, si se anuncia algo esperado, como en "mañana el sol se elevará sobre el horizonte" o "el gobierno lamenta la muerte del Papa", la información es cercana a 0. Pero si la probabilidad es más baja, como en "mañana temblará en Bogotá" o "Colombia invade a Estados Unidos", el contenido de información es mayor.

Desde el punto de vista de la eficiencia, es razonable esperar que la transmisión de eventos probables requiera menos tiempo (un menor costo promedio) que la transmisión de eventos sorpresivos. Un ejemplo es el código Morse, cuyos signos más comunes (más probables) se representan en general con menos caracteres y con más caracteres los que menos se usan (cuadro 1). Los caracteres son rayas y puntos, o sus representaciones binarias. Algunas vocales, especialmente la e, y otras letras muy usadas en inglés, tienen códigos cortos de dos o tres caracteres. Hay un mensaje en Morse que contiene una gran cantidad de información, compuesto por pocos símbolos fáciles de reproducir: el pedido de auxilio, SOS, que se transmite como una secuencia de tres puntos, tres rayas y tres puntos:

...———... ...———... ...———... ...———...

Cuadro 1
Código Morse

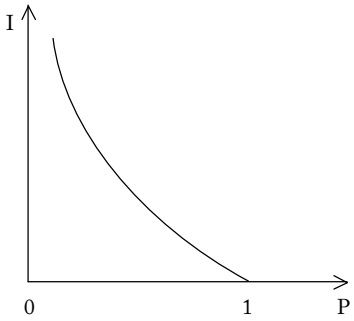
A	.–	M	––	Y	–.––	6	–....
B	–...	N	–.	Z	––..	7	––...
C	–.–.	O	–––	Ä	.–.–	8	–––..
D	–..	P	.––.	Ö	–––.	9	––––.
E	.	Q	––.–	Ü	..––	.	.–.–.–
F	..–.	R	.–.	Ch	––––	,	––..––
G	––.	S	...	0	–––––	?	..––..
H	....	T	–	1	.––––	!	..__.
I	..	U	..–	2	..–––	:	–––...
J	.–––	V	...–	3	...––	“	.–..–.
K	–.–	W	.––	4	....–	‘	.––––.
L	.–..	X	–..–	5	.....	=	–...–

MÉTRICA DE LA INFORMACIÓN

Los primeros intentos de enmarcar el tema de la información en una teoría formal surgieron en los años veinte, con el trabajo de Hartley (1928), y luego con el trabajo de Shannon (1948), quien subrayó los conceptos probabilísticos en el tratamiento de la información y de su transmisión[1].

Ya se mencionó que la cantidad de información de un mensaje se relaciona con la probabilidad de ocurrencia del evento reportado: a mayor sorpresa más información. En términos formales, sea I la cantidad de información y P la probabilidad del evento. I es entonces una función *f* del inverso de P, I = f(1/P), como se ilustra en la gráfica 1.

Gráfica 1



Además, es razonable suponer que la información I es una cantidad positiva, continua en P, y que la información contenida en dos eventos independientes es la suma de la información individual. Se puede mostrar que la única función que cumple estas propiedades es la logarítmica (ver, p. ej., Chen y Alajaji, 2005, cap. 2). Así, la cantidad de información de un mensaje I se define como:

$$I = \log(1/P) = -\log P$$

donde I es igual a 0 cuando P es igual a 1 (el evento ya se conocía o la noticia se esperaba con total certeza), y muy grande cuando P tiende a 0 (para eventos sorprendentes, como los milagros).

La probabilidad P depende en gran parte de la frecuencia con la que el evento en cuestión haya ocurrido anteriormente y de la manera como va cambiando la percepción de la gente. En consecuencia, la cantidad de información I puede cambiar en el tiempo. Por ejemplo,

[1] Para mayores detalles de la formalización de estos conceptos, ver Lathi (1974), Thomas (1975), Krippendorff (1986), Brémaud (1988), Chen y Alajaji (2005), y Gray (2009). Para una extensión del concepto de información a documentos públicos y noticias, ver Montenegro (1995).

la probabilidad de que mañana el sol no se eleve sobre el horizonte es casi 0, pero si mañana no se eleva, la probabilidad de que no se eleve pasado mañana aumenta, a la vez que disminuye el contenido de información de ese mismo mensaje en el futuro.

ENFOQUE TÉCNICO

Se puede llegar a una definición técnica del contenido de información similar a la que se obtiene de manera intuitiva.

El problema es similar a un problema de optimización sujeto a restricciones impuestas por la tecnología disponible y la naturaleza, como el ruido. Los mensajes se originan en una fuente y deben transmitirse a través de un medio o canal hasta llegar a su destino o receptor. Esto se debe hacer minimizando la distorsión que puedan causar el medio de transmisión y el ruido y, a la vez, maximizando el número de mensajes enviados por unidad de tiempo, o costo. Es, en resumen, un problema de eficiencia. Por tanto, la cantidad de información de un mensaje es proporcional al costo de transmisión.

Desde el punto de vista técnico, la teoría de la información se concentra en la transmisión binaria, es decir, en la transmisión de unos y ceros, pulsos y no pulsos, rayas y puntos, etc. Un bit (*binary unit*) es algo que puede tomar uno de esos dos estados.

Cuando solo hay dos mensajes posibles, por ejemplo, {llovió, no llovió}, basta enviar un uno o un cero. Uno si llovió y cero si no llovió, o viceversa; en todo caso, esta asignación, llamada codificación, debe ser conocida por el receptor. Este uno o cero enviado es un bit. En otras palabras, para enviar un mensaje con dos posibilidades solo es necesario un bit. Pero si se trata de más posibilidades, por ejemplo, {no llovió, llovió poco, llovió el promedio y llovió mucho}, necesitamos dos bits de información para transmitir estos cuatro mensajes, un uno o un cero seguido por otro uno o cero. Así formamos los códigos 00, 01, 11 o 10, a los que podemos asignar los cuatro mensajes.

Si fueran ocho mensajes necesitaríamos tres bits de información para formar los ocho códigos, 000, 001, 011, 111, 110, 100, 101, 010, necesarios para distinguir los ocho mensajes. En resumen, para enviar 2^i mensajes se requieren $I = \log_2 2^i = i$ bits de información.

Cuando los mensajes tienen diferentes probabilidades de ocurrencia pueden existir esquemas distintos para asignar códigos con el mismo número de dígitos binarios a cada mensaje de modo que la transmisión sea más eficiente, es decir, que los mensajes se puedan transmitir con menos de $\log_2 2^i$ bits en promedio. En esta formulación

es conveniente expresar el contenido de información directamente en términos de probabilidades.

Supongamos que podemos enviar uno de 8 mensajes: A, B, C, D, E, F, G, H, los cuales tienen diferente probabilidad de ocurrencia. En principio, como ya se señaló, se podrían codificar con tres dígitos binarios asignándoles los códigos 000, 001, 011, 111, 110, 100, 101, 010, lo que requeriría $\log 2^3 = 3$. Pero como los mensajes tienen diferentes probabilidades se puede encontrar un esquema más eficiente (que requiera menos bits). Siguiendo un ejemplo que se encuentra en Touretzky (2004), supongamos que los 8 mensajes tienen probabilidades, 1/2, 1/4, 1/8, 1/16, 1/32, 1/64, 1/128, 1/128, que suman 1. En vez de codificar con tres dígitos cada mensaje, usamos un código de longitud variable. Si enviamos el mensaje A transmitimos 0, si enviamos B transmitimos 10; así, 010 representa AB, y así sucesivamente, como se muestra a continuación:

A	0
B	10
C	110
D	1110
E	11110
F	111110
G	1111110
H	1111111

En la codificación anterior no se usan, por ejemplo, los códigos 1 para B o 00 para C porque en una secuencia de transmisión no se sabría dónde empiezan y dónde terminan los códigos de cada mensaje, donde el 0 indica el fin de un mensaje (excepto para H, el cual se reconoce fácilmente que está compuesto por siete unos). La transmisión de cada mensaje individual requiere tantos bits como dígitos binarios tenga su código, pero el valor esperado del esquema de codificación es menor que 3; según esta codificación la transmisión requeriría, en promedio:

$$\frac{1}{2}+\frac{2}{4}+\frac{3}{8}+\frac{4}{16}+\frac{5}{32}+\frac{6}{64}+\frac{7}{128}+\frac{7}{128} = 1{,}98 \text{ bits},$$

menos de 3 bits. El resultado anterior debe entenderse en el contexto de un gran número de repeticiones, de un promedio o valor esperado, pues la transmisión de un solo mensaje requiere un costo proporcional a la longitud de su código. Es en la repetición donde se obtienen ahorros en términos del costo promedio por mensaje. En otras palabras,

en el ejemplo anterior la transmisión de 1.000 mensajes requeriría 1.980 bits en promedio, en vez de 3.000. El promedio anterior es la entropía que se discute a continuación.

ENTROPÍA

El resultado anterior se formaliza partiendo de un conjunto de mensajes independientes $\{m_1, m_2,..., m_n\}$ llamados alfabeto, con probabilidades $\{P_1, P_2,..., P_n\}$. Los mensajes pueden ser las letras A, B,…, H del ejemplo anterior o los símbolos del alfabeto Morse.

Si se envía una serie N de estos mensajes, donde N es grande, habrá aproximadamente P_1N mensajes m_1, P_2N mensajes m_2 y así sucesivamente. Todas las secuencias S tendrán la misma probabilidad de ocurrencia porque contendrán el mismo número de mensajes m_i, probabilidad que será expresada por el producto de las probabilidades de repetir cada m_i, donde $i = 1, 2,..., n$:

$$P(S) = (P_1)^{P_1 N}(P_2)^{P_2 N}\cdots(P_n)^{P_n N}$$

Como ejemplo, supongamos los eventos {llovió, no llovió} con probabilidades {0,1, 0,9}. Si N es grande la probabilidad de cualquier secuencia será la misma: $(0,1)^{0,1N}(0,9)^{0,9N}$, aunque el orden de envío de los mensajes cambia de una serie otra.

Retomando la definición de información como el logaritmo del inverso de la probabilidad P(S), podemos escribir la información de la secuencia S como:

$$I(S) = \log_2 \frac{1}{P(S)} = \log_2 \left(\frac{1}{P_1}\right)^{P_1 N}\left(\frac{1}{P_2}\right)^{P_2 N} \cdots \left(\frac{1}{P_n}\right)^{P_n N} = N\sum_{i=1}^{n} P_i \log_2 \frac{1}{P_i}$$

Ahora definimos:

$$H = \frac{I(S)}{N} = \sum_{i=1}^{n} P_i \log_2 \frac{1}{P_i} = -\sum_{i=1}^{n} P_i \log_2 P_i$$

como el contenido promedio de información por mensaje, o sea la esperanza matemática de cada mensaje. También se puede interpretar como el contenido promedio de la incertidumbre asociada a la fuente de mensajes. La variable H fue introducida por Shannon (1948), quien la llamó entropía, tomando prestado el nombre de un concepto de la mecánica estadística que usa la misma fórmula. La entropía representa el mínimo número de bits (costo) que se requiere en promedio para transmitir un mensaje.

En la fórmula de H se considera que si la probabilidad de uno de los mensajes es 0 su contribución a la entropía también es 0, lo cual se justifica matemáticamente porque:

$$\lim_{P \to 0+} P \log P = 0$$

En general, en términos de una variable x y su probabilidad, definimos la entropía como el valor esperado del contenido de información:

$$H(X) = E[-\log_2 P(x)] = -\sum_{x \in X} P(x) \log_2 P(x)$$

La entropía H(X) es mayor cuanto más uniforme (más equiprobable) es la distribución de X, y llega a un máximo cuando todos los elementos de X tienen la misma probabilidad. Si X puede tomar un número L de valores o estados posibles, se puede demostrar que la entropía está acotada como sigue:

$$0 \leq H(X) \leq \log L$$

expresión que es igual a 0 si P(x)=1, lo que implica que X es determinística y tiene 0 incertidumbre, e igual a logL si P(x)=1/L para todo x (ver Massey, 1998, cap. 1). Es decir, la máxima entropía se alcanza cuando todos los estados posibles tienen la misma probabilidad de ocurrencia.

Como se señaló antes, las unidades se denominan bits cuando la base del logaritmo es 2. Pero se puede utilizar una base diferente; para el logaritmo natural las unidades de información se denominan nats. En adelante se usa la base 2 salvo que se indique lo contrario.

LA ENTROPÍA EN LA FÍSICA

El concepto de entropía se desarrolló en la segunda mitad del siglo XIX, primero en termodinámica por Rudolf Clausius, quien acuñó el término, y luego en mecánica estadística por Ludwig Boltzmann, quien la presentó en términos probabilísticos. En física, la entropía es una variable de estado, como la temperatura, el volumen, la presión o la energía interna, que describe el estado del sistema en un momento dado, en contraste con variables como velocidad, posición o masa que pueden describir las condiciones individuales de las moléculas que conforman el sistema.

En termodinámica, y de manera informal, la entropía tiene que ver con el hecho de que no toda la energía que entra en un proceso (p. ej., el accionar de un motor) se aprovecha para realizar trabajo ya

que parte se pierde en forma de calor o ruido. Si bien la energía se conserva, parte se transforma en formas menos útiles para realizar trabajo. La fórmula derivada en termodinámica relaciona el cambio en la entropía con la transferencia de calor Q realizada a una temperatura T, y se expresa como:

$$\Delta H = \frac{Q}{T}$$

Se puede demostrar que esta formulación es compatible con la definición de entropía en mecánica estadística:

$$H = k \log \Omega$$

en apariencia más cercana a la fórmula de Shannon que a la fórmula termodinámica. En la formulación de mecánica estadística k es la constante de Boltzmann y Ω es el número de estados que puede tomar un sistema. Por ejemplo, si se lanza un dado, Ω es 6; si se lanzan dos y se registra el resultado de las caras, Ω es 36. En física, un sistema puede contener billones de moléculas, cada una de las cuales puede ser descrita con la ayuda de varias coordenadas de manera que Ω alcanza fácilmente el orden de billones de billones. Si suponemos estados equiprobables y notamos que, en esencia, el inverso de Ω es la probabilidad de cada uno de ellos ($P = 1/\Omega$) y hacemos $k = 1$, nos acercamos a la noción de entropía de la información:

$$H = \log \Omega = \frac{1}{\Omega} \sum \log \Omega = -\sum \frac{1}{\Omega} \log \frac{1}{\Omega}$$

En física, la entropía es una medida del desorden o del caos del sistema. En teoría de la información, es una medida de la incertidumbre asociada a la fuente de mensajes. No hay consenso sobre si hay o no hay relación entre ambos tipos de entropía, excepto que tienen fórmulas similares.

La primera ley de la termodinámica dice que la energía se conserva. La segunda ley de la termodinámica, o ley de la entropía, implica que, si bien la energía se conserva, el tipo de energía utilizable para hacer trabajo disminuye o al menos no aumenta en un sistema cerrado. La entropía en un sistema cerrado mide la cantidad de energía inutilizable. La segunda ley implica que el calor fluye del cuerpo con más temperatura al cuerpo con menos temperatura y no al contrario. Hoy es menos popular la definición de entropía como aumento del desorden o caos y más popular la definición de dispersión espontánea de energía o como medida de procesos irreversibles.

ENTROPÍA MULTIVARIADA

La definición de entropía puede extenderse a vectores de variables. Para el vector X, Y, caracterizado por la distribución de probabilidad conjunta P(x, y), la entropía conjunta se escribe:

$$H(X,Y) = -\sum_{x\in X,\, y\in Y} P(x,y)\log P(x,y)$$

Para variables independientes, cuya distribución conjunta se puede expresar como P(x, y) = P(x)P(y), la entropía es aditiva:

$$H(X,Y) = H(X) + H(Y)$$

De manera similar, se puede definir la entropía condicional como una cantidad que indica el comportamiento de la información o incertidumbre de una variable cuando se conoce otra:

$$H(X/Y) = -\sum_{x\in X,\, y\in Y} P(x,y)\log P(x/y)$$

Recordemos que estamos escribiendo el valor esperado de P(x/y), que es una función de x y y, de modo que la distribución de probabilidad apropiada es la distribución conjunta de x y y, P(x, y). Si X puede tomar un número de L valores o estados posibles, se puede demostrar que:

$$0 \le H(X/Y) \le \log L$$

igual a 0 si P(x/y) = 1 para algún x e igual a logL si P(x/y) = 1/L para todo x.

Haciendo uso de P(x, y) = P(x)P(y/x) = P(y)P(x/y), la entropía conjunta puede relacionarse con la entropía condicional a través de:

$$H(X,Y) = H(X) + H(Y/X) = H(Y) + H(X/Y)$$

Se puede demostrar que H(X/Y) ≤ H(X), cumpliendo la igualdad cuando X y Y son independientes (ver Chen y Alajaji, 2005); el conocimiento de Y da información adicional que puede reducir la incertidumbre sobre X (pero nunca aumentarla). En el extremo, H(X/Y) = 0 indicaría que X se torna determinística luego de conocer Y. De lo anterior se deduce que H(X, Y) ≤ H(X) + H(Y), cumpliendo la igualdad cuando X y Y son independientes.

El concepto de información mutua se define como:

$$I(X;Y) = H(X) - H(X/Y) = H(Y) - H(Y/X) = \sum_{x\in X}\sum_{y\in Y} P(x,y)\log\frac{P(x,y)}{P(x)P(y)}$$

donde el último término de la derecha carece de signo negativo porque equivale a H(X/Y) – H(X). Nótese que la expresión es simétrica, es decir, I(X; Y) = I(Y; X). La información mutua refleja la información que una variable proporciona acerca de la otra; mide la dependencia entre X y Y. En general:

$$0 \leq I(X; Y) \leq \min[H(X), H(Y)]$$

que es igual a 0 cuando X y Y son independientes e igual a min[H(X), H(Y)] cuando X y Y se relacionan determinísticamente. Otra formulación es:

$$I(X; Y) = H(X) + H(Y) - H(X, Y)$$

denotando la información mutua como la intersección de H(X) y H(Y).

Otro concepto muy útil es el de entropía relativa de Kullback-Leibler, que mide la divergencia o distancia entre dos distribuciones de probabilidad definidas sobre los mismos valores de x:

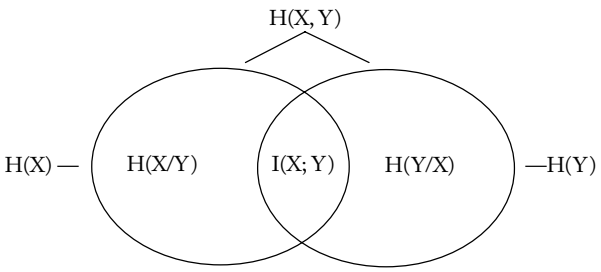
$$D_{KL}(P,Q) = E\left(\log\frac{P(x)}{Q(x)}\right) = \sum_{x\in X} P(x)\log\frac{P(x)}{Q(x)}$$

Se puede demostrar que $D_{KL}(P, Q) \geq 0$ e igual a 0 si ambas distribuciones son iguales para todo x (ver, p. ej., Massey, 1998, cap. 1). Por otro lado, la entropía relativa de Kullback-Leibler no es simétrica, $D_{KL}(P, Q) \neq D_{KL}(Q, P)$; si P(x) es diferente de 0 mientras algún Q(x) es 0, $D_{KL}(P, Q) = \infty$ pero $D_{KL}(Q, P)$ no. Nótese que la información mutua puede expresarse como:

$$I(X; Y) = D_{KL}(P(x, y), (P(x)P(y))$$

La gráfica 2, tomada de Chen y Alajaji (2005, 37), muestra un diagrama de Venn que ilustra la interrelación de estos conceptos.

Gráfica 2



También existen versiones de entropía para variables continuas:

$$H(X) = -\int f_X(x)\log f_X(x)dx$$

$$H(X,Y) = -\iint f_{X,Y}(x,y)\log f_{X,Y}(x,y)dxdy$$

$$H(X/Y) = H(X,Y) - H(Y) = -\iint f_{X,Y}(x,y)\log\frac{f_{X,Y}(x,y)}{f_Y(y)}dx\,dy$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y) = \iint f_{X,Y}(x,y)\log\frac{f_{X,Y}(x,y)}{f_X(x)\,f_Y(y)}dx\,dy$$

las cuales no necesariamente son positivas porque las funciones de densidad continuas pueden tomar valores mayores de 1.

APUESTAS

Una de las primeras aplicaciones de estos conceptos de información y entropía a un campo distinto de la codificación y transmisión de información se encuentra en un artículo de Kelly (1956) sobre las apuestas con información privilegiada. Kelly trata varios casos. En todos, un apostador obtiene información sobre el resultado de una apuesta a través de un canal de comunicación contaminado por ruido que puede inducir errores. El apostador recibe la información antes que el público en general.

El primer caso trata de dos posibles resultados, es decir, un resultado binario donde el mensaje transmitido tiene una probabilidad p de ser transmitido con error y una probabilidad 1 – p de ser transmitido sin error. Si el apostador apuesta todo su capital (a todo o nada) en cada una de las N rondas, maximiza su valor esperado:

$$EV_N = [2(1-p)]^N V_0$$

donde V_0 es el capital inicial y V_N el final. Sin embargo, si N es grande el apostador perderá todo en algún momento con una probabilidad igual a 1. Si en vez de apostar todo apuesta una fracción ℓ de su capital tenemos:

$$V_N = (1+\ell)^W (1-\ell)^L V_0$$

donde W es el número de veces que gana y L es el número de veces que pierde en las N rondas. La tasa de ganancia exponencial será:

$$G = \lim_{N\to\infty} \frac{1}{N} \log \frac{V_N}{V_0} = \lim_{N\to\infty} \left[\frac{W}{N} \log(1+\ell) + \frac{L}{N} \log(1-\ell) \right] = (1-p)\log(1+\ell) + p\log(1-\ell)$$

donde G es la ganancia por período en $2^{GN} = V_N/V_0$ (si en vez del logaritmo en base 2 usamos el logaritmo natural, la expresión es $e^{GN} = V_N/V_0$). Esta expresión se maximiza con respecto a ℓ derivando e igualando a 0, y se obtiene:

$$\frac{1-p}{1+\ell} - \frac{p}{1-\ell} = 0, \text{ de donde } \ell = 1-2p$$

Remplazando en G, obtenemos:

$$G_{max} = (1-p)\log 2(1-p) + p\log 2p = 1 + (1-p)\log(1-p) + p\log p$$

que depende de la entropía de la transmisión. Nótese que, por ser una entropía binaria, la cantidad $(1-p)\log(1-p)+p\log p$ varía entre 0 y -1, con lo cual $0 \le G_{max} \le 1$. Kelly muestra que, en el largo plazo, un apostador que escoja la fracción ℓ así obtenida, superará a otro que invierta una fracción diferente.

Generalizando, en el segundo caso Kelly supone varios mensajes de entrada excluyentes, s, y no necesariamente equiprobables. La información privilegiada –en el sentido de que al apostador se le transmite el resultado (de una carrera de caballos, p. ej.) antes que el público en general lo conozca– puede estar distorsionada por ruido o porque el apostador no confía plenamente en su fuente. Por ello se involucran probabilidades condicionales en el cálculo. Kelly supone que el apostador mantiene invertido todo su capital:

$$\sum_s a(s/r) = 1$$

donde a(s/r) es la proporción del capital que apuesta a s después de recibir el mensaje r. El capital del apostador evoluciona así:

$$V_N = \prod_{r,s} \left[a(s/r)\alpha_s \right]^W V_0$$

donde W es el número de veces que se transmite s y se recibe r, y α_s son relaciones de probabilidades (*odds ratio*), esto es, el número de veces que se multiplica el valor apostado si ocurre s.

Al final de la primera ronda, V_1 será igual a V_0 multiplicado por $a(s/r)\alpha_s$ (con los datos del s que haya ganado en esa ronda); al final de la segunda ronda, V_2 será igual a V_1 multiplicado por $a(s/r)\alpha_s$ (con los datos del s que haya ganado en la segunda ronda), y así sucesivamente.

Para el s ganador, el capital aumenta en las rondas donde $a(s/r)\alpha_s > 1$ y disminuye en las rondas donde $a(s/r)\alpha_s < 1$.

Tomando logaritmos y un N grande, obtenemos la tasa de ganancia exponencial:

$$\log \frac{V_N}{V_0} = \sum_{r,s} W \log \alpha_s a(s/r)$$

$$G = \lim_{N\to\infty} \frac{1}{N} \log \frac{V_N}{V_0} = \sum_{r,s} P(s,r) \log \alpha_s a(s/r)$$

$$= \sum_{s} P(s) \log \alpha_s + \sum_{r,s} P(s,r) \log a(s/r)$$

donde P(s,r) es la probabilidad conjunta de s y r, y P(s) es la probabilidad de s.

Suponiendo que las apuestas sean justas (*fair odds*), esto es $\alpha_s = 1/P(s)$, tenemos:

$$G = \sum_{s} P(s) \log \frac{1}{P(s)} + \sum_{r,s} P(s,r) \log a(s/r) = H(X) + \sum_{r,s} P(s,r) \log a(s/r)$$

donde $H(X) = \sum_{s} P(s) \log \frac{1}{P(s)}$ es la entropía de la fuente.

Dado un resultado recibido r, la ganancia G se maximiza escogiendo a(s/r). La expresión del lagrangiano $\mathcal{L}$, sujeto a $\sum_{s} a(s/r) = 1$, es:

$$\mathcal{L} = \sum_{s} P(s,r) \log a(s/r) - \lambda (1 - \sum_{s} a(s/r))$$

cuyas condiciones de primer orden son:

$$\frac{\partial \mathcal{L}}{\partial a(s/r)} = \frac{P(s,r)}{a(s/r)} - \lambda = 0$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = 1 - \sum_{s} a(s/r) = 0$$

De la primera condición se obtiene:

$$\sum_{s} a(s/r) = \frac{1}{\lambda} \sum_{s} P(s,r)$$

Y sustituyéndola en la segunda se llega a:

$$\lambda = \sum_{s} P(s,r)$$

Así, la primera se transforma en:

$$\frac{P(s,r)}{a(s/r)} - \sum_{j} P(j/r) = 0$$

Finalmente, despejando $a(s/r)$ obtenemos:

$$a(s/r) = \frac{P(s,r)}{\sum_{j} P(j,r)} = \frac{P(s,r)}{P(r)} = P(s/r)$$

lo que implica que el apostador maximiza su ganancia escogiendo cada $a(s/r)$ de manera proporcional a la probabilidad de s dado el mensaje recibido r. El valor máximo de la tasa de ganancia será:

$$G_{max} = H(X) + \sum_{r,s} P(s,r) \log P(s/r) = H(X) - H(X/Y)$$

donde $H(X) - H(X/Y) = I(X;Y)$ es la información mutua antes descrita, que mide la dependencia entre variables y tiende a 0 cuando hay independencia total y es mayor que 0 según el grado de dependencia.

Aun si $\alpha_s \neq 1/P(s)$, lo que quiere decir que las probabilidades no corresponden a las apuestas hechas por el público (*unfair odds*), Kelley demuestra que la ganancia G se máxima escogiendo $a(s/r) = P(s/r)$, es decir, que el apostador ignora las probabilidades implícitas en las apuestas que hace el público α_s,

$$G_{max} = \sum_{r,s} P(s,r) \log P(s/r) + \sum_{s} P(s) \log \alpha(s) = -H(X/Y) + H(\alpha)$$

donde $H(\alpha) = \sum_{s} P(s) \log \alpha(s)$.

En otros contextos, las apuestas públicas (*odds*) podrían corresponder a las predicciones de los analistas del mercado bursátil, mientras que las probabilidades condicionales $P(s/r)$ podrían ser las que surgen del análisis propio del inversionista, incluyendo información privilegiada.

ALGUNAS APLICACIONES ECONÓMICAS

En economía, el concepto de entropía se aplica en dos áreas: aquellas que tienen semejanzas con la termodinámica, como la degradación

de las habilidades productivas, y aquellas que se refieren al contenido de la información y su utilización en el análisis de datos.

El creciente interés en la combinación de conceptos económicos y físicos se aprecia en publicaciones como la revista *Physica A*, cuya sección permanente titulada "Econophysics" incluye estudios que analizan variables económicas desde la óptica de la mecánica estadística y usando conceptos teóricos de la física: entropía, leyes de potencia, movimiento browniano, etc.

Georgescu-Roegen (1971) examina los límites que la entropía impone al crecimiento económico; su analogía del reloj de arena ilustra cómo un sistema cerrado, sin intercambio de energía con el exterior, pierde capacidad para hacer trabajo productivo. La arena que va cayendo pierde su energía potencial o habilidad para realizar trabajo. Puede argumentarse que nuestro sistema no es cerrado ya que recibe energía del sol. Sin embargo, esta energía fluye a una tasa finita, como la arena que cae por la restricción en la mitad del reloj. El autor subraya que este concepto de restricción en la tasa de transmisión de energía está ausente en los modelos neoclásicos, empezando por el diagrama tradicional del flujo circular de la producción y la distribución.

Samuelson (1986) y Smith y Foley (2002) encuentran analogías entre la teoría económica neoclásica y las leyes de la termodinámica. En un estudio de la sostenibilidad del crecimiento económico, McMahon y Mrozek (1997) aceptan que la teoría neoclásica incluye el efecto de la primera ley de la termodinámica sobre la conservación de la materia y la energía, pero consideran necesario complementarla con el concepto de entropía o segunda ley de la termodinámica, según la cual los procesos son irreversibles en el tiempo. Esto implica que si bien la energía y la materia se conservan en un sistema cerrado, como el sistema solar, la calidad y el tipo de insumos nunca se recuperan a partir del producto, degradando paulatinamente el acervo de recursos.

Por otra parte, la economía de la información estudia la información y el conocimiento como bienes económicos; su calidad, su mercado, y cómo afectan las decisiones de los agentes. La economía de la información tiene orígenes en los trabajos de Hayek (1945), donde se diferencia la información que se mantiene en las torres de marfil y la información económicamente útil, y cómo opera esta última a través del sistema de mercado. En este sentido, difiere del enfoque de la teoría de la información que se expone en este artículo, porque la información como inverso de la probabilidad y la entropía como información promedio son conceptos más adecuados para medir la

cantidad de información que la calidad, el uso o el tipo de información, y más apropiados para estudios estadísticos. Con un enfoque más semejante al de Hayek, Domenech (1989) presenta una visión filosófica del papel de la información en la sociedad, y cómo se restringe y se filtra según el sistema político-económico.

Aplicaciones en finanzas

Se han hecho intentos para adaptar los conceptos asociados a la medición del contenido de información al estudio de temas financieros y bursátiles. La hipótesis de que los mercados responden al nivel de incertidumbre o al elemento sorpresa tiene analogía con la entropía.

Darbellay y Wuertz (2000) recuerdan que las series de rendimientos financieros no siguen distribuciones normales sino distribuciones con colas gruesas, de mayor probabilidad, y que es difícil aplicar el teorema del límite central a estas series por la dependencia serial de los rendimientos (p. ej., en sus cuadrados, como en los modelos ARCH). El concepto de entropía es útil para estudiar esta dependencia porque no se limita a la dependencia lineal. Estos autores estudian, por separado, dos series de rendimientos financieros: la tasa de cambio dólar-marco entre octubre de 1992 y mayo de 1997 registrada cada 30 minutos y el índice diario Dow Jones en el periodo 1901-1998. Usando diversos rezagos τ en cada serie, observan pocas diferencias entre la estimación de la información mutua $I(r(t); r(t-\tau))$ y la estimación de la información mutua del valor absoluto $I(|r(t)|;|r(t-\tau)|)$, y concluyen que prácticamente toda la información de los rendimientos en t se encuentra en los valores absolutos de los rendimientos pasados y no en sus signos, excepto para rezagos muy cortos. Este ejercicio es similar a la estimación de la función de autocorrelación, ACF, pero difiere en que recoge todo tipo de dependencia, no solo la lineal. Algunos autores se refieren a la información mutua acerca de la misma serie como función de autoinformación.

En un estudio similar, Maasoumi y Racine (2002) aplicaron la entropía para encontrar dependencias no lineales de los rendimientos de los precios bursátiles y sus predicciones. Por su parte, Chen (2002) discute la similitud entre los conceptos de teoría de la información y el valor económico de la información con respecto a los mercados. La cantidad de información recibida por un inversionista depende de su entrenamiento, y entenderla es un proceso de aprendizaje que toma tiempo, por lo cual, aunque la información esté públicamente disponible, su difusión, evaluación y uso apropiado no son inmediatos como supone la hipótesis del mercado eficiente.

La medida de desigualdad de Theil

La entropía se presta como medida de heterogeneidad. Theil (1967) la adaptó para construir su conocido índice de desigualdad del ingreso:

$$T = \sum_i y_i \log N y_i$$

Para ello, en la fórmula de la entropía remplazó la probabilidad por y_i, la proporción del ingreso total correspondiente al individuo i. Con esta notación, la entropía del ingreso es:

$$H(y) = -\sum_i y_i \log y_i$$

La máxima equidad se alcanza cuando todos los y_i son iguales, digamos $y_i = 1/N$, donde N es el número de individuos de la sociedad, que al remplazar en la fórmula anterior da:

$$H(y) = -\sum_i \frac{1}{N} \log \frac{1}{N} = \log N$$

La medida de desigualdad de Theil, T, es la entropía máxima menos la entropía de la distribución del ingreso:

$$T = \log N - H(y) = \log N + \sum_i y_i \log y_i$$

Y puesto que $\sum_i y_i = 1$ se llega a fórmula usual:

$$T = \sum_i y_i \log N + \sum_i y_i \log y_i = \sum_i y_i \log N y_i$$

El método de estimación de máxima entropía

Sea $\{x_i, y_i\}$ una muestra de tamaño N, donde x_i es un vector de variables explicativas y y_i la variable dependiente. El objetivo es estimar la función de probabilidad condicional P(y/x). La distribución empírica es:

$$\hat{P}(x, y) = \frac{1}{N} \text{ (\# de veces que x, y ocurre en la muestra)}$$

Supongamos además que se conocen las condiciones o características específicas de algunos de los N eventos. Por ejemplo, si es frecuente que y aumente cuando x_1 es mayor que x_2. Estas condiciones, compa-

tibles con la muestra y útiles para la estimación, se pueden expresar por medio de variables *dummy* R(x, y). Por ejemplo, R(x, y) = 1 si y y x cumplen la condición, y R(x, y) = 0 si no la cumplen. A partir de la muestra se construye el valor esperado de R(x, y):

$$\hat{P}(R) = \sum_{x,y} \hat{P}(x,y)R(x,y)$$

el cual hacemos igual a la probabilidad que debe arrojar el modelo, es decir, imponemos la restricción:

$$\hat{P}(R) = \sum_{x,y} \hat{P}(x,y)R(x,y) = \sum_{x,y} \hat{P}(x)P(y/x)R(x,y) = P(R)$$

donde $\hat{P}(x)$ es la distribución de x en la muestra.

El criterio de máxima entropía elije, sujeto a las restricciones, la distribución condicional P(y/x) más cercana a la distribución uniforme, porque implementa el criterio básico del método: no suponer lo que no se sabe. Recordando que la entropía es proporcional a la uniformidad de una distribución (cuanto más uniforme mayor la entropía), maximizamos la entropía condicional encontrando P(y/x) en la siguiente función objetivo:

$$-\sum_{x,y} \hat{P}(x)P(y/x)\log P(y/x) + \lambda[\sum_{x,y} \hat{P}(x,y)R(x,y) - \sum_{x,y} \hat{P}(x)P(y/x)R(x,y)] + \gamma[\sum_{y} P(y/x) - 1]$$

donde λ y γ son multiplicadores de Lagrange.

Una comparación de los métodos de máxima entropía y de mínimos cuadrados ordinarios se encuentra en Eruygur (2005). En Colombia, Morley et al. (1998) utilizan este método para estimar la movilidad de ingresos.

Otras aplicaciones econométricas

El conocido criterio de información de Akaike para identificar modelos se basa en el concepto de información de Kullbak-Leibler[2]. Para una densidad *f* que representa el modelo verdadero y una densidad g del modelo que se va a probar, la entropía relativa de Kullback-Leibler puede escribirse como:

$$D_{KL} = \sum_{x \in X} f(x)\log\frac{f(x)}{g(x)} = E_f\left[\log f(x) - \log g(x)\right]$$

[2] Ver Akaike (1974) o Anderson et al. (2000).

Suponiendo que el modelo que representa la realidad es desconocido pero constante ante los modelos alternativos, la expresión anterior se simplifica:

$$D_{KL} = C - E_f[\log g(x)]$$

donde C es una constante. La idea es entonces probar modelos g para minimizar la pérdida con respecto al verdadero f. Al minimizar el valor esperado de la pérdida de información dado por D_{KL}, se deriva la expresión del criterio de Akaike, $-\frac{2\ell}{n} + \frac{2k}{n}$, donde ℓ es la verosimilitud logarítmica maximizada, k el número de parámetros estimados y n el número de observaciones utilizadas en la estimación (ver Burnham, 2004).

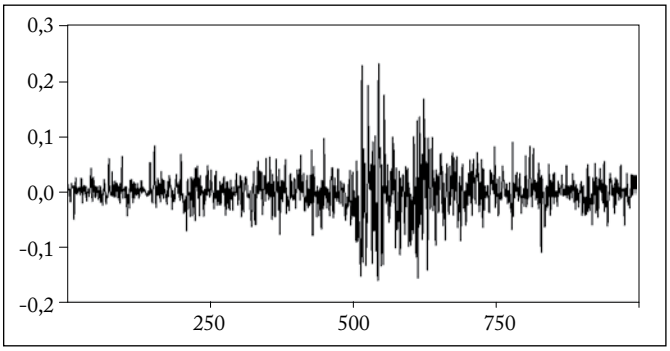
El concepto de Kullback-Leibler también se ha aplicado a la interpretación del R^2 como contenido de información, es decir, de la incertidumbre explicada por el modelo estimado. Cameron y Windmeijer (1995) proponen un seudo R^2 para modelos de regresión no lineales basado en la reducción de la incertidumbre generada por la inclusión de variables explicativas, donde dicha reducción es medida por la divergencia de Kullback-Leibler.

Ejercicio econométrico

El siguiente ejercicio ilustra el uso de las variables relacionadas con el contenido de la información en la predicción de los precios bursátiles. Para ellos se toman las variaciones porcentuales del precio de cierre de la acción de Alcoa, la empresa productora de aluminio que es el primer componente de los 30 que forman el índice Dow Jones. La serie se muestra en la gráfica 3. En el ejercicio se hace una regresión de la variación de los precios contra sus rezagos, sus rezagos al cuadrado y los rezagos de una variable derivada del contenido de información de la serie de variaciones porcentuales del precio.

La razón para incluir el cuadrado de la variación del precio es que la variable de información tiene un comportamiento similar al del cuadrado de los precios (a mayor varianza mayor contenido de información) y que, en presencia de estos cuadrados, podría ser redundante. Así, la inclusión del cuadrado de los precios en las regresiones hace más exigente la prueba de la variable de información.

Gráfica 3
Alcoa – variación porcentual del precio diario de cierre
(26 de septiembre de 2006-14 de septiembre de 2010)

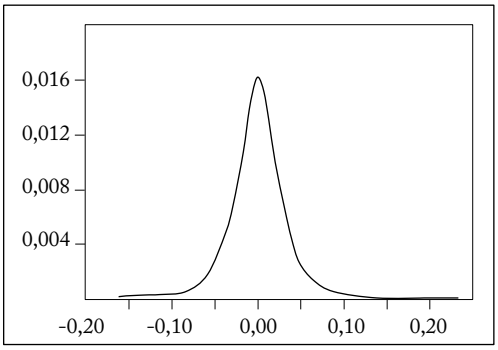


Como las variaciones de los precios pueden ser positivas o negativas, mientras que la información es positiva, la variable de información que se utiliza es el producto de la información por la variación porcentual del precio de la acción. Si d_t es la variación porcentual del precio de la acción, la variable de información es:

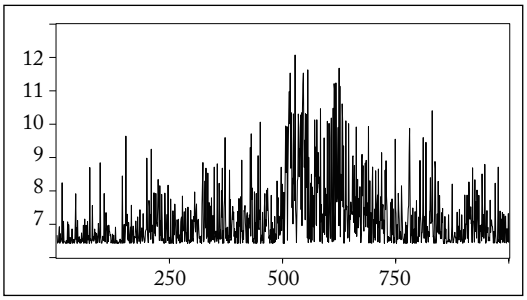
$d_t(-\log P(d_t))$

La gráfica 4 muestra la estimación de la distribución de probabilidad de la serie de la variación porcentual del precio de Alcoa, la gráfica 5 el contenido de información $-\log P(d_t)$ y la gráfica 6 la variable de información $d_t(-\log P(d_t))$.

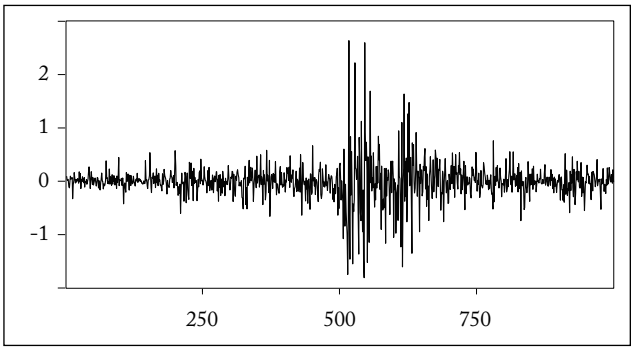
Gráfica 4
Estimación Kernel de la distribución de probabilidad de la
variación porcentual del precio de Alcoa



Gráfica 5
Alcoa – contenido de información de la variación del precio vs. tiempo



Gráfica 6
Variable de información



Los datos se tomaron de la página financiera de Yahoo y corresponden a observaciones diarias del 25 de septiembre de 2006 al 14 de septiembre de 2010. La regresión se representa así:

$$d_t = c + \sum_{s=1}^{40} \alpha_s d_{t-s} + \rho_s d_{t-s}^2 + \gamma_s d_{t-s}(-\log P(d_{t-s}) + \varepsilon_t$$

La significancia conjunta de los coeficientes de los 40 rezagos de las tres variables explicativas se probó con el estadístico F. Además, se probó la significancia de los 40 coeficientes de la variable de información. Los resultados fueron los siguientes:

Acción	R^2	adj R^2	F-stat	p-F-stat	P-Wald (información)	P-Wald (ruido)
Alcoa	0,262	0,157	2,483	0,000	0,000	0,927

La regresión es estadísticamente significativa, como indican las columnas 4 y 5. Por su parte, la variable de información es estadísticamente significativa para predecir el precio de la acción, como muestra la columna 6. A manera de contraste, la columna 7 muestra el resultado de la repetición de la regresión añadiendo 40 rezagos de una serie de ruido blanco como variable explicativa, los cuales no resultaron significativos.

CONCLUSIÓN

El concepto de información, aunque tiene diferentes interpretaciones según la disciplina, se puede interpretar en términos de costo-beneficio. Transmitir, recibir y entender textos o mensajes requiere un esfuerzo que se justifica si el beneficio esperado es superior al costo. El beneficio depende del nivel de información contenido en el mensaje y este, a su vez, se relaciona con el grado de sorpresa del mensaje transmitido. Cuanto más baja es la probabilidad de ocurrencia del evento comunicado, mayor es su contenido de información, y viceversa; de modo que existe una relación inversa entre probabilidad e información. Esta relación probabilística tiene muchas aplicaciones. En economía, por ejemplo, para modelar el desarrollo y sus limitaciones; en econometría, para diseñar índices y métodos de estimación, y en el campo financiero, para explicar el comportamiento de los mercados bursátiles.

REFERENCIAS BIBLIOGRÁFICAS

1. Akaike, H. "A New Look at the Statistical Model Identification", *Transactions on Automatic Control* 19, 1974, pp. 716-723.
2. Anderson D., K. Burnham y W. Thompson. "Null hypothesis testing: Problems, prevalence and an alternative", *Journal of Wildlife Management* 64, 4, 2000, pp. 912-923.
3. Brémaud, P. *An introduction to probabilistic modeling*, Springer-Verlag, 1988.
4. Burnham, K. "Multimodel inference: Understanding AIC and BIC in model selection", paper presented at the Amsterdam Workshop on Model Selection, August 27-29, 2004.
5. Cameron C. y F. Windmeijer. "An R squared measure of goodness of fit for some common nonlinear regression models", Department of Economics, University of California, Davis, 1995.
6. Chen, J. "Information, entropy and evolutionary finance", Working Paper, School of Business, University of Northern British Columbia, 2002.
7. Chen P. y F. Alajaji. "Lecture notes on information theory", Department of Communications Engineering, National Chiao Tung University, and Department of Mathematics & Statistics, Queen's University, 2005.

8. Darbellay G. y D. Wuertz. "The entropy as a tool for analyzing statistical dependence in financial time series", *Physics A* 287, 2000, pp. 429-439.
9. Domenech, A. *De la ética a la política*, Barcelona, Editorial Crítica, 1989.
10. Eruygur, O. "Generalized maximum entropy (GME) estimator: Formulation and a Monte Carlo study", *MPRA Paper* 12459, 2005.
11. Georgescu-R. N. *The entropy law and the economic process*, Cambridge, Harvard University Press, 1971.
12. Gray, R. *Entropy and information theory*, New York, Springer-Verlag, 2009.
13. Hartley, R. "Transmission of information", *Bell System Technical Journal* 7, 1928, pp. 535-563.
14. Hayek, F. "The use of knowledge in society", *American Economic Review* 35, 4, 1945, pp. 519-530.
15. Krippendorff, K. *Information theory: Structural models for qualitative data*, Beverly Hills, Sage Publications, 1986.
16. Kelly, J. L. "A new interpretation of information rate", *Bell System Technical Journal* 35, 1956, pp. 917-926.
17. Lathi, B. P. *Sistemas de comunicación*, México, Limusa, 1974.
18. Maasoumi, E. y J. Racine. "Entropy and predictability of stock market returns", *Journal of Econometrics* 107, 2002, pp. 291-312.
19. Massey, J. "Applied digital information theory", Lecture Notes, ETH (Instituto Federal de Tecnología), Zürich, 1998.
20. McMahon G. y J. Mrozek. "Economics, entropy and sustainability", *Hydrological Sciences Journal* 4, 24, 1997, pp. 501-512.
21. Montenegro, Á. "El contenido de información en documentos y mensajes", *Documentos CEDE* 95-06, 1995.
22. Morley, S., S. Robinson y R. Harris. "Estimating income mobility in Colombia using maximum entropy econometrics", *TMD Discussion Paper* 26, 1998.
23. Samuelson, P. *Collected economics papers*, vol. 5, Cambridge, MIT Press, 1986.
24. Shannon, C. "A mathematical theory of communication", *The Bell System Technical Journal* 27, 1948, pp. 379-423 y 623-656.
25. Smith, D. y D. Foley. "Is utility theory so different from thermodynamics?", *SFI Working Paper* 02-04-016, 2002.
26. Theil, H. *Economics and information theory*, Amsterdam, North Holland, 1967.
27. Thomas, J. "Information, communication, noise and interference", D. Fink, ed., *Electronics engineers handbook*, New York, McGraw-Hill, 1975.
28. Touretzky, D. S. "Basics of information theory", 2004, [http://www.cs.cmu.edu/~dst/Tutorials/Info-Theory/], consulta en diciembre de 2009.